AN INTERACTIVE MEDIA DECISION SUPPORT SYSTEM

David Ness and Christopher R. Sprague*

595-72

March 1972

AN INTERACTIVE MEDIA DECISION SUPPORT SYSTEM

David Ness and Christopher R. Sprague*

595-72

March 1972

To be presented as "Interactive Media Planning" at IEEE Computer Society Conference, Tokyo, Japan, June 1972.

*Visiting Associate Professor of Industry
Wharton School of Finance and Commerce
University of Pennsylvania

ABSTRACT

Media planning consists of (1) establishing a desired market; (2) finding media (in particular, periodicals and television shows) which reach that market; and (3) allocating advertising funds to these media. This paper discusses the design, implementation and operation of an interactive, terminal-based computer system which manages a large data base and provides models that aid the advertising decision maker in solving this problem.

AN INTERACTIVE MEDIA DECISION SUPPORT SYSTEM

David Ness and Christopher R. Sprague

The problem of allocating advertising dollars to media is an important one to three distinct groups of people:

a. advertisers - who have money for advertising and want to reach potential consumers of their product,

b. media - who have advertising space which they want to sell to advertisers and

c. agencies - who are principally involved in trying to get (a) and (b) together in an efficient and effective way.

The system which is described in this paper contributes to the solution of these problems by bringing together some large syndicated data bases and a number of pre-existing and newly defined models through the medium of an interactive, time-shared computer system. The system which is described in this paper is the principal product of Interactive Market Systems, Inc.[1] which has been delivering this service to their clients since early 1970. In this paper we first discuss the problem mentioned above and then turn to a brief chronological history of the implementation of the system. We will then describe some of the general characteristics of that system and close with some observationsabout the way and nature that the system has grown since it first began to be used by customers two years ago.

[1] Interactive Market Systems, Inc., 360 Lexington Avenue, New York City.

Statement of Problem

In any mass market the advertising media serve a vital and important information dissemination function which allows potential consumers of a product and the producers of a product to get together in the marketplace. In such a market, since the producers are relatively few and the consumers relatively many, it is the problem of the producer to reach the consumer to inform him of the characteristics of his product. In order to accomplish this objective advertising media exist as an effective channel of communication. Before he can place an ad, however, the advertiser must answer a number of questions:

a. who are the potential consumers for the product?

b. what are their media habits (for example, what magazines do they read and what television shows do they watch)?

c. howcan we effectively use the media to reach the consumers?

d. what tactics should be used (for example, what content should be put in the ad, what characteristics of the product should be emphasized, and what should the frequency of advertisements be)?

The system which we describe here addresses, for the most part, only questions (a), (b) and (c) above. Although some models have been suggested in an attempt to deal with parts of the tactical problem (see, for example, Lodish and Little[2]), these have not found wide acceptance in the industry and it is thought at this time that such questions are perhaps best resolved by the advertising agency. Therefore, the system described here does not concentrate on this problem but rather emphasizes the problem of identifying potential consumers and effectively relating them to their media habits in

[2] Lodish, L. M. and Little, J. D. C., "MEDIAC-An On-Line Media Planning System," Fall Joint Computer Conference, 1968.

such a way as to determine the overall strategy of an advertising campaign. Tactical questions are left for resolution by the advertisers and their agencies.

How Can the Problem Be Addressed?

Since this problem has been around for a substantial period of time, several contributions have been made in an attempt to effect a reasonable solution. For example, a number of syndicated data bases are available for advertisers, media and agencies to purchase. One such data base is the well known "Simmons Sample" collected annually by W. R. Simmons and Associates Research, Inc.[3] We decided to focus our initial development efforts on this data base.

The sample consists of an annual survey of more than 15,000 representative consumers. Each of these consumers is administered a lengthy questionnaire by an interviewer and an attempt is made to elicit a profile of the consumer on four separate kinds of dimensions:

a. demographic characteristics (income level, number of people in family, location of residence, etc.),

b. consumption habits (information about consumption of widely used products by brand with a number of detailed questions about heavily advertised products in particular),

c. a description of media habits consisting of a two-week television diary and two separate measurements of magazine readership and

d. psychographic information about the consumer's image of himself and his ideal image.

[3] W. R. Simmons and Associates Research, Inc., 235 East 42nd Street, New York City.

This data in its disaggregated form represents some 10 to 15 thousand bits of information from each consumer. From this data Simmons and other similar companies prepare a large number of cross-tabulated summary reports which might, for example, consist of information about magazine readership breaking down the population into nonsmokers, smokers of filtered cigarettes, smokers of menthol cigarettes and smokers of nonfiltered cigarettes. These pretabulated reports often prove useful to the advertisers, agencies and the media but since they necessarily represent only a small fraction of the information available in the total sample, it has been customary for users of the information to request that specially prepared tab reports be generated to their specifications. This normally involves processing requests through a batch-oriented computer system and this necessarily entails delays in obtaining response to the requests of from eight to 72 hours depending upon the nature of the problems involved in getting a batch run made.
It was to help cut down on these delays that we became interested in attempting to develop a time shared facility for accessing this large data base, operating on the general hypothesis that if we could cut down on the delays involved in the process, then the users of the information might be prone to use it more extensively and effectively.

Initial System Specifications and Implementation

We began to work on the system in earnest in late 1969. At that point in time the problems involved in building a system were divided into three distinct components:

a. the development of a user language,

b. the development of a program organization and strategy and

c. the development of a strategy for storing the rather large quantity of data present in the Simmons data base.

The first of these tasks, the development of the user language, was important for two reasons. First, users of the pre-existing batch processing system had not themselves directly formulated the requests for their reports. Their requests were typically telephoned in to a company representative who would then reformulate the request in terms that the computer system which generated tabulated reports would understand. This representative thus functioned as a technical intermediary between the users of the data and the computer system. If the on-line strategy were to prove to be at all effective, then it was necessary that the users of the data be able to forumuate their own requests at their own office site and thus not expect to have a technical intermediary present to instruct them in how to formulate their requests. Second, if it were to prove to be too difficult for the user to operate the system, then he probably would not do so. This, of course, would mean the system, although perhaps technically sound, would not prove to be useful. For this reason, a substantial amount of time in the early design was invested in developing a simple and straightforward user-oriented language which would allow him to specify reports in a clear and concise way and provide for relatively simple and direct interaction with the system.

The question of program organization was also important for two reasons. First, since computer time would be a major component in delivering information to the consumer, it was necessary to develop a strategy where these costs could be kept to a reasonable minimum. Second, it was recognized that if the system were successful, then it would be necessary to adapt it and allow it to grow over time without necessitating tremendous changes both in the pre-existing program and to the user interface which accessed these programs.

For this reason it was necessary to develop an organization strategy which was flexible enough to be able to grow substantially over time without incurring costs of rewriting a large portion of the system.

Decisions about storage strategy were also important. They were complicated first by the size of the data base. Early in the design process, a committment was made to deliver any report which could logically be constructed from the data in the data base to any one who requested it. This necessitated keeping all of the data in the data base around in disaggregated form, and made it impossible to reduce the large quantity of data present there simply by aggregating over some of the dimensions. This suggested that it would be fruitful to try and compact the data as much as possible without losing any of the data itself. To some extent it is necessary in such circumstances to face a tradeoff between computation costs and storage costs in that any compaction scheme usually involved incurring costs both for the compaction of the data and its later expansion. In this case, however, it was felt that most of the data expansion could be performed by the computer while other data was being retrieved and that thus the net effect on computer costs should be to reduce them.

Further aspects of the data should be mentioned. First, all of the data in the data base was potentially binary in character in that all real and integer valued data items were recorded in discrete range scales at interview time. Thus information about income, for example, is not maintained by storing the actual income of the respondent but rather by simply recording the fact that his income is in the range of $6000-8000. Second, and perhaps most important of all, as the data is originally collected

and stored on the tapes delivered by the data supplier, it is organized by question within respondent. Thus all of the data about respondent 1 appears in the first "card" of the data base and all of the data about respondent 2 in the second "card," etc. Answering questions in an on-line system with the data stored in this form would obviously be very expensive both in time and money, as any request for retrieval would involve scanning through the entire data base. Thus, the alternative of inversion was a necessity in this case. Costs are incurred by inverting the data file into one organized by respondent within question, but since it is only necessary to perform this function once each year when the new data is collected, this cost can be amortized over all of the retrievals which are performed during the year. This proves a cost effective tradeoff, since now to answer any specific question it is only necessary to access those parts of the data base which refer to that question, thus allowing a great saving both in elapsed time and money.

At this point it seemed logical to divide the system efforts into two distinct parts. First, a development of an interactive system which would carry on a dialogue with the user, retrieve requested information and print the desired reports. The second part of the system (the so-called periodic system) would be run only when it was necessary to build a new data base from basic data. This part of the system would also have the responsibility to allow the operations personnel to develop new directories through which the users would access the information. It was necessary to introduce several levels of directories in order to provide flexibility in the user language and also to adapt the system to different characteristics of each year's sample as, for example, questions are often added or deleted or the

responses to the questions recorded in different places in the basic data set. Later a third part of the system developed, namely that associated with auxilliary functions such as customer billing and allowing the customer to make estimates of the cost of a given run prior to actually making that run. As will be mentioned below it proved to be useful to tie some of these auxilliary system quite directly into the interactive system.

Growth of Models

Shortly after the basic system began to function it became clear that while users of the system found it useful to be able to retrieve information from this data base very quickly, it did not contribute very directly to the solution of any of the problems we outlined above in the first section of this paper. We found, for example, that the results of a run with the system were often used as input to one of the many well known (in the advertising business) models which attempted to relate this basic information to quantities of more direct interest to the manager. An example of one such model is the "Metheringham reach and frequency model."[4] Simply put, the Metheringham model allows one to take data about the actual readership patterns of a population for several different media and bring this together with a proposed advertising schedule stated in terms of number of insertions in that series of media in an attempt to get an estimate of the portion of the population which would see one or more of the advertisements (the so-called net reach of the given schedule). Since many of the users of the data base did Metheringham reach and frequency analyses, it was thought appropriate to incorporate this procedure as part of the system. Because of the general structure of the system this proved to be an easy task. We will now try to

[4] Metheringham, R. A., "Measuring the Net Cumulative Coverage of a Print Campaign," Journal of Advertising Research, December 1964.

illuminate this point by giving an overall view of the basic design of the system and follow that up with an indication of how this new model was implanted in it. We will later note that several other models have subsequently been incorporated into the system.

One of the constraints that we operated under was a time-sharing service which charged us in proportion to the space-time product of our program.[5] Such a charging scheme strongly suggested that we attempt to keep our core images small and caused us early in the design to adopt which we called a "phase structure" where each part of the calculation would be performed in a phase and when that phase terminated operation, it would be thrown out of core and a new phase started to continue the calculation. All communications between phases, due to limitations of the time-sharing system, had to be through files of disk storage as it was not possible to leave information in core memory from one phase to the next without writing one's own supervisor. While we first expected that this charging scheme would interfere with our programming "style," we found it to be a valuable discipline in that the program phase represents a natural unit for both program execution and conceptual documentation. This discipline caused us to develop several different programming strategies which will be mentioned briefly below and are discussed in greater detail in Ness, Sprague and Moulton.[6]

This concept of phase structure also suggested the opportunity to do what we call direct posting. With this method of program organization each

[5] The actual product was number of seconds of execution times number of thousands of words of core memory.

[6] Ness, David, Sprague, C. R. and Moulton, G. A., "On the Implementation of Sophisticated Interactive Systems," Sloan School of Management Working Paper 506-71, MIT, 1971.

phase, as it decides that it has something to communicate to another phase, simply writes that into a file which will be read only by that other phase, Thus, any modifications in the system which take place in between two phases which are communicating need not and will not interfere with the form of communication between those two phases. We found this notion tremendously useful as our initial decision of what constituted phases began to undergo transformation due to computational realities. For example, at one point in time the entire process of compiling a user request for information into an actual data retrieval was handled by a single phase. This proved, given the charging scheme mentioned above, to be quite an ineffective use of resources and we were able to cut costs substantially by splitting apart the operation into several distinct phases. All of this could be done without, for example, interfering with the communication between the initial input reader and the final report writer, which consisted of such things as report titles, row stubs, column headers and information about the actual constituents of the final report.

User Language

The system, when used in its basic information retrieval mode, requests that the user provide four kinds of input to specify his final report:

a. things describing the report itself (title, stubs, headers, whether row percentages are desired, etc.),

b. the definition of the population that the user desires to scan,

c. which portions of the population fall in the categories defined by rows of the report, and

d. which portions of the population fall in the categories defined by columns in the report.

A typical row or column specification might be a statement like:

```
MIDINC MEN: MEN.AND.INC IN(6000:8000);
```

or

```
PLAYBOY:AUD(PLA);
```

Each of these lines consists of a stub or header (which precedes the colon) and a Boolean specification of the characteristics of people who it is desired fall into the column or row being specified. In the first example above we have a stub suggesting middle income men (MIDINC MEN) and then we define this as having the properties from the data base of being male (this is obviously just a yes or no) and having an income in the range $6000-8000 (another item directly obtainable from the original survey). In the second example we are asking for a specification of the average audience of a magazine (this is a very common question, indeed). Since all magazines are surveyed for two time periods, the average audience is the number who indicated reading during the first time period (X) plus the number who indicated reading during the second time period (Y) divided by two. In order to avoid making the user type a mathematical equation like:

```
(RXPLA+RYPLA)/2
```

the concept of the audience function (AUD) was introduced. The system translates this request into a form which allows the information retrieval to proceed directly and then it automatically generates the averaging function.

The specification of a population is fully as general as the specification of a row or column, the only difference being that results for each population are printed in separate reports and thus the report title is assumed to describe the population. Therefore, there is no stub or header associated with this kind of item.

In general, any specification can be an arbitrary Boolean condition on any constituents in the data base. In addition, the user is able to refer to an item in the data base in any one of three different languages:

a. by a mnemonic name (for example, MEN),

b. by an @ name which relates to the form in which the data is encoded by the original questionnaire (for example, @01091 represents a one punch in column nine of card one of the original survey and this happens to be where the answer to the question Man? is located) or

c. by a # quantity which refers directly to the way the data is actually stored in the data base (#1 is block one of the system's data storage, #2 is block two of the data storage, etc.).

Almost all users make their requests in mnemonic or @ form while systems operations personnel often find it useful to be able to refer to the actual structure of data by # for system reliability checking and maintenance.

Logging Information

In the early stages of system implementation it was decided to have each major system operation log its effect in a disk file which could be looked at if difficulties arose. Information written into this log file consists of such quantities as

a. job identification,

b. time of day,

c. elapsed amount of computer time,

d. cost up to this point in the calculation and

e. indication of what activity is going on at this point in time.

This information proved to be immensely useful during the debugging of the system. When errors arose and crashes occurred, it was possible to look at the log and find out the most recent major activity of the system and thus get a good notion of the cause of the difficulties. It later proved to be the case that it was possible by the addition of a few extra logging commands to actually bill customers directly from the system log file. Thus in one of the first auxilliary programs written to enable the system to produce customer bills, all that was necessary was to scan the system log file and the bill could be produced.

Perhaps most important of all, by looking at and summarizing the systems log operations personnel could easily locate those parts of the overall system which proved to be incurring largest costs. By identifying these high cost stages of the process, we could then focus our attention almost exclusively on these elements in an attempt to get costs reduced. Thus, for example, while we have always had the option to optimize the actual referencing of the disk during data retrieval, it has proved that this is not a very substantial cost element in the system and as a result this has never been done. On the other hand, the portion of the system devoted to tabulating the final report proved to be a major cost component and by focusing efforts on this area we were able to achieve cost reductions of in excess of 70%. A decision to include a system log proved to be tremendously important with respect to the overall computational efficiency of the process.

Data Retrieval and Tabulation

After the user's input has been compiled into a series of requests for retrieval and Boolean manipulation of basic items in the data base, we

proceed to perform these retrievals and calculations and produce a new data file which consists of a series of binary vectors representing whether a consumer is (or is not) in the population which is being surveyed, whether he is within the definition of row one of the report, row two of the report . . . and within the definition of column one of the report, column two of the report, etc. This data file is then handed over to a tabulation phase which simply rereads each vector and actually counts the individuals in the appropriate cells of the final report. Since it proves to be statistically effective to oversample large consumers (a stratified sampling), it is also necessary to accumulate weights of each consumer during this final tabulation. These weights correct for the over- or under-representation of given parts of the population and allow us to produce a final report consisting of either actual counts from the sample or projected counts on the base of the American adult population when the over- or under-representation is taken into account. In most cases users are interested in the latter, but we always maintain the former information so that we are able to flag cells in the final report which are based on thin (less than 60 actual people) or very thin (less than 30 actual people) samples from the data. This gives the user of the information some indication of the reliability of his estimates. The tabulation phase is also sent a signal from one of the first phases of the system to tell it which phase it should call next. Thus, if we are doing a reach and frequency analysis, one of the reach and frequency phases will be called into operation, while if we are doing a simple information retrieval, the standard report writer can be called. As we will see in a moment, there has been a substantial growth in

the number of different options and this philosophy has left us flexibility with respect to adding new ones.

System Growth

Throughout this paper we have emphasized the importance of allowing the system to grow in ways which do not interfere with its basic operation. Our ability to do this has been tested several times. First, when the Metheringham reach and frequency procedure was added, it became clear that users did not like to be required to specify the large number of individual data items which must be retrieved from the data base in order to apply this model. Given the phase structure of the system it was therefore important that we were able to write a new user interface to handle this kind of request and incorporate this into the system by simply having it produce as its output the input that the user would otherwise have had to specify directly to the retrieval system. For example, to evaluate a schedule involving insertions of advertisements in Time, Life, Playboy and Ladies Home Journal, it is only necessary that the user specify (a) that he wants to do a Metheringham analysis and (b) that the media he is concerned with are TIM, LIF, PLA and LAD. The Metheringham input analyzer will then generate a request for a rather large cross-tab (in this case, 8 by 8) consisting of the audience of each of the media in each of the time periods surveyed. This request can then be handed over to the standard system with an indication that once tabulation has been performed, the system should exit to the Metheringham reach and frequency model, not to the standard report writer. When the R and F model again seizes control, the user is able to type in a specific schedule (three insertions in Time and two

in Playboy, for example) and obtain his desired reach and frequency report. This vastly simplifies the user's input process and makes the system more useful to him.

This philosophy was further tested by the addition of an optimization package. After watching customers use the Metheringham reach and frequency, it was noticed that they very often were involved in a process of doing a rough heuristic optimization of a schedule according to some relatively simple budget constraints. Since the model involves satisfies all of the properties necessary to be able to apply branch and bound techniques, it was decided that perhaps the system could do this more efficiently than the user. An optimization package was added to the system which requests additional information from him (for example, the cost of ads in a magazine and the total budget available). Then this optimization program uses the Metheringham R and F model to evaluate various possibilities and gradually iterates to an optimum schedule subject to the given constraints. An interesting comment at this point is that initial surveys of the customers to ascertain whether they would use such a procedure indicated that it would not prove to be a viable product. Nevertheless, once it became available and users learned what it could do for them, it has proved to be a widely (and we think productively) used tool.

Since the initial reach and frequency and optimization procedures were added, several other models have been incorporated. Two alternative methods of calculating reach and frequency (both of which produce more reliable results but involve somewhat greater computation costs) and an optional flexible report generator have been added. Further, the system has grown

to incorporate data bases other than Simmons. For example, at the moment the syndicated data base of Brand Rating Index[7] is currently available in the domestic system and several other data bases are in the process of being "brought up." Further, a version of the system is currently operating in London, England with the Simmons data (for American subsidiaries in London and for English companies with large American markets) and two large English data bases are available. Thus the system seems to be growing in terms of number of models available, number of data bases available and geographically all at the same time.

System Documentation

The documentation for the system is kept on-line. This proved to be important for several reasons:

a. one always knows where the definitive version of the documentation can be found (in the on-line system),

b. we have had more success in getting people to document on-line than we have had with previous, conventional off-line (i.e. write it on paper) modes and

c. being able to machine process the documentation allows us to perform some functions such as programmer control and job assignment somewhat more easily than would otherwise have been the case.

On-line documentation has also proved to be reasonably cheap, costing us on the order of (not, of course, including time of the documenter which would be involved anyway) 10 to 20 cents per program documented. We find that we are able to get a fully sorted list of our documentation file prepared for less than $2 and thus cost has not proved to be a substantial

[7] Brand Rating Index, Inc., New York City.

impediment to using the system. This documentation system is described in greater detail in Ness.[8]

Acknowledgements

The authors are indebted to Leon Liebman, President of IMS and Assistant Professor, Wharton School, University of Pennsylvania for his advice and encouragement. We also appreciate the efforts of Richard Makely and Paul Albrecht and others at IMS for their patience and understanding in communicating the desires and frustrations of actual users of the system to us.

[8] David N. Ness, "On Line Documentation System," Sloan School Working Paper, MIT (forthcoming).